

---

# Bayesian Linear Regression

Prof. Nicholas Zabarar  
Center for Informatics and Computational Science  
<https://cics.nd.edu/>  
University of Notre Dame  
Notre Dame, Indiana, USA

Email: [nzabarar@gmail.com](mailto:nzabarar@gmail.com)  
URL: <https://www.zabarar.com/>

January 31, 2019



# Contents

---

- [Bayesian inference in linear regression when  \$\sigma^2\$  is unknown](#), [Zellner's g-Prior](#), [Uninformative \(Semi-Conjugate\) Prior](#), [Evidence Approximation](#)
- [Bayesian model comparison](#), [Model Complexity](#), [The evidence approximation for our regression example](#), [Another example of computing model evidence](#)
- [Limitations of fixed basis functions](#)
- [Laplace approximation](#), [BIC criterion](#), [Another Regression example and implementation of model selection](#)

Following closely:

- [Chris Bishops' PRML book](#), Chapter 3
- Kevin Murphy, [Machine Learning: A probabilistic Perspective](#), Chapter 7
- [Regression using parametric discriminative models in pmtk3](#) (run [TutRegr.m](#) in [Pmtk3](#))



# Bayesian Inference when $\sigma^2$ is Unknown

- Let us extend the previous results for linear regression assuming now that  $\sigma^2$  is unknown.
- Assume a likelihood of the form:\*

$$p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{y} | \mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N) = \frac{1}{(2\pi)^{N/2}} (\sigma^2)^{-N/2} \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})}{2\sigma^2}\right)$$

- A *conjugate prior* has the following form:

$$p(\mathbf{w}, \sigma^2) = \mathcal{N}\mathcal{IG}(\mathbf{w}, \sigma^2 | \mathbf{w}_0, \mathbf{V}_0, a_0, b_0) \triangleq \mathcal{N}(\mathbf{w} | \mathbf{w}_0, \sigma^2 \mathbf{V}_0) \mathcal{IG}(\sigma^2 | a_0, b_0) \\ \frac{b_0^{a_0}}{(2\pi)^{D/2} |\mathbf{V}_0|^{1/2} \Gamma(a_0)} (\sigma^2)^{-(a_0 + D/2 + 1)} \exp\left(-\frac{(\mathbf{w} - \mathbf{w}_0)^T \mathbf{V}_0^{-1} (\mathbf{w} - \mathbf{w}_0) + 2b_0}{2\sigma^2}\right)$$

- The posterior is now derived as:

$$p(\mathbf{w}, \sigma^2 | \mathcal{D}) = \frac{b_0^{a_0}}{(2\pi)^{(N+D)/2} |\mathbf{V}_0|^{1/2} \Gamma(a_0)} (\sigma^2)^{-(a_0 + (D+N)/2 + 1)} \exp\left(-\frac{(\mathbf{w} - \mathbf{w}_0)^T \mathbf{V}_0^{-1} (\mathbf{w} - \mathbf{w}_0) + (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + 2b_0}{2\sigma^2}\right)$$

❖ In this lecture, the response is denoted as  $\mathbf{y}$ , the dimensionality of  $\mathbf{w}$  as  $D$  and the design matrix as  $\mathbf{X}$ .



# Bayesian Inference when $\sigma^2$ is Unknown

$$p(\mathbf{w}, \sigma^2 | \mathcal{D}) = \frac{b_0^{a_0}}{(2\pi)^{(N+D)/2} |\mathbf{V}_0|^{1/2} \Gamma(a_0)} (\sigma^2)^{-(a_0+(D+N)/2+1)} \exp\left(-\frac{(\mathbf{w} - \mathbf{w}_0)^T \mathbf{V}_0^{-1} (\mathbf{w} - \mathbf{w}_0) + (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + 2b_0}{2\sigma^2}\right)$$

□ Let us define the following:

$$\mathbf{V}_N = (\mathbf{V}_0^{-1} + \mathbf{X}^T \mathbf{X})^{-1}, \quad \mathbf{w}_N = \mathbf{V}_N (\mathbf{V}_0^{-1} \mathbf{w}_0 + \mathbf{X}^T \mathbf{y})$$

$$a_N = a_0 + N/2, \quad b_N = b_0 + \frac{1}{2} (\mathbf{w}_0^T \mathbf{V}_0^{-1} \mathbf{w}_0 + \mathbf{y}^T \mathbf{y} - \mathbf{w}_N^T \mathbf{V}_N^{-1} \mathbf{w}_N)$$

□ With these definitions, one with simple algebra can show:

$$p(\mathbf{w}, \sigma^2 | \mathcal{D}) \propto (\sigma^2)^{-(a_N+D/2+1)} \exp\left(-\frac{(\mathbf{w} - \mathbf{w}_0)^T \mathbf{V}_0^{-1} (\mathbf{w} - \mathbf{w}_0) + (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + 2b_N - \mathbf{w}_0^T \mathbf{V}_0^{-1} \mathbf{w}_0 - \mathbf{y}^T \mathbf{y} + \mathbf{w}_N^T \mathbf{V}_N^{-1} \mathbf{w}_N}{2\sigma^2}\right)$$

$$\propto (\sigma^2)^{-(a_N+D/2+1)} \exp\left(-\frac{(\mathbf{w} - \mathbf{w}_N)^T \mathbf{V}_N^{-1} (\mathbf{w} - \mathbf{w}_N) + 2b_N}{2\sigma^2}\right)$$

$$p(\mathbf{w}, \sigma^2 | \mathcal{D}) = \mathcal{NIG}(\mathbf{w}, \sigma^2 | \mathbf{w}_N, \mathbf{V}_N, a_N, b_N) \triangleq \mathcal{N}(\mathbf{w} | \mathbf{w}_N, \sigma^2 \mathbf{V}_N) \mathcal{IG}(\sigma^2 | a_N, b_N)$$

□ The posterior marginals can now be derived explicitly:

$$p(\sigma^2 | \mathcal{D}) = \mathcal{IG}(\sigma^2 | a_N, b_N)$$

$$p(\mathbf{w} | \mathcal{D}) = \mathcal{F}_D\left(\mathbf{w}_N, \frac{b_N}{a_N} \mathbf{V}_N, 2a_N\right) \propto \left[1 + \frac{(\mathbf{w} - \mathbf{w}_N)^T \mathbf{V}_N^{-1} (\mathbf{w} - \mathbf{w}_N)}{2b_N}\right]^{-\frac{2a_N+D}{2}}$$



# Posterior Marginals

$$p(\mathbf{w} | \mathcal{D}) \propto \int_0^\infty (\sigma^2)^{-(a_N + D/2 + 1)} \exp\left(-\frac{(\mathbf{w} - \mathbf{w}_N)^T \mathbf{V}_N^{-1} (\mathbf{w} - \mathbf{w}_N) + 2b_N}{2\sigma^2}\right) d\sigma^2 \propto \left[1 + \frac{(\mathbf{w} - \mathbf{w}_N)^T \mathbf{V}_N^{-1} (\mathbf{w} - \mathbf{w}_N)}{2b_N}\right]^{-\frac{2a_N + D}{2}}$$

□ The marginal posterior can be directly written as:

$$p(\mathbf{w} | \mathcal{D}) = \mathcal{J}_D\left(\mathbf{w}_N, \frac{b_N}{a_N} \mathbf{V}_N, 2a_N\right) \propto \left[1 + \frac{(\mathbf{w} - \mathbf{w}_N)^T \mathbf{V}_N^{-1} (\mathbf{w} - \mathbf{w}_N)}{2b_N}\right]^{-\frac{2a_N + D}{2}}$$

❖ To compute the integral above, simply set  $\lambda = \sigma^{-2}$ ,  $d\sigma^2 = -\lambda^{-2} d\lambda$  and use the normalizing factor of the Gamma distribution  $\int_0^\infty \lambda^{a-1} e^{-b\lambda} d\lambda = \Gamma(a) b^{-a} \sim b^{-a}$ .



# Posterior Predictive Distribution

□ Consider the posterior predictive for  $m$  new test inputs:

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathcal{D}) \propto \iint \frac{1}{(2\pi)^{m/2}} (\sigma^2)^{-m/2} \exp\left(-\frac{(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{w})^T (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{w})}{2\sigma^2}\right) (\sigma^2)^{-(a_N + \underbrace{D/2+1}_{\leftarrow})} \exp\left(-\frac{(\mathbf{w} - \mathbf{w}_N)^T \mathbf{V}_N^{-1} (\mathbf{w} - \mathbf{w}_N) + 2b_N}{2\sigma^2}\right) d\mathbf{w} d\sigma^2$$

□ As a first step, let us integrate in  $\mathbf{w}$  by writing:

These terms cancel out from the integration in  $\mathbf{w}$

$$\begin{aligned} & (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{w})^T (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{w}) + (\mathbf{w} - \mathbf{w}_N)^T \mathbf{V}_N^{-1} (\mathbf{w} - \mathbf{w}_N) + 2b_N = \\ & \left( \mathbf{w} - (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} + \mathbf{V}_N^{-1})^{-1} (\tilde{\mathbf{X}}^T \tilde{\mathbf{y}} + \mathbf{V}_N^{-1} \mathbf{w}_N) \right)^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} + \mathbf{V}_N^{-1}) \left( \mathbf{w} - (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} + \mathbf{V}_N^{-1})^{-1} (\tilde{\mathbf{X}}^T \tilde{\mathbf{y}} + \mathbf{V}_N^{-1} \mathbf{w}_N) \right) \\ & - (\tilde{\mathbf{X}}^T \tilde{\mathbf{y}} + \mathbf{V}_N^{-1} \mathbf{w}_N)^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} + \mathbf{V}_N^{-1})^{-T} (\tilde{\mathbf{X}}^T \tilde{\mathbf{y}} + \mathbf{V}_N^{-1} \mathbf{w}_N) + \mathbf{w}_N^T \mathbf{V}_N^{-1} \mathbf{w}_N + \tilde{\mathbf{y}}^T \tilde{\mathbf{y}} + 2b_N \end{aligned}$$

□ Let us denote the last term in the Eq. above as

$$2\beta = -(\tilde{\mathbf{X}}^T \tilde{\mathbf{y}} + \mathbf{V}_N^{-1} \mathbf{w}_N)^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} + \mathbf{V}_N^{-1})^{-T} (\tilde{\mathbf{X}}^T \tilde{\mathbf{y}} + \mathbf{V}_N^{-1} \mathbf{w}_N) + \mathbf{w}_N^T \mathbf{V}_N^{-1} \mathbf{w}_N + \tilde{\mathbf{y}}^T \tilde{\mathbf{y}} + 2b_N$$



# Posterior Predictive Distribution

□ The posterior predictive

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathcal{D})$$

$$\propto \iint \frac{1}{(2\pi)^{m/2}} (\sigma^2)^{-m/2} \exp\left(-\frac{(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{w})^T (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{w})}{2\sigma^2}\right) (\sigma^2)^{-(a_N+D/2+1)} \exp\left(-\frac{(\mathbf{w} - \mathbf{w}_N)^T \mathbf{V}_N^{-1} (\mathbf{w} - \mathbf{w}_N) + 2b_N}{2\sigma^2}\right) d\mathbf{w} d\sigma^2$$

is now simplified using  $\lambda = 1/\sigma^2$  and recalling the normalization of the [Gamma distribution](#):

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathcal{D}) \propto \int (\lambda)^{m/2+a_N-1} \exp(-\beta\lambda) d\lambda \sim \beta^{-(m/2+a_N)}$$

□ Substituting  $\beta$  and by comparing the two Eqs. one can verify:

$$\begin{aligned} p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathcal{D}) &\propto \left( -(\tilde{\mathbf{X}}^T \tilde{\mathbf{y}} + \mathbf{V}_N^{-1} \mathbf{w}_N)^T (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} + \mathbf{V}_N^{-1})^{-T} (\tilde{\mathbf{X}}^T \tilde{\mathbf{y}} + \mathbf{V}_N^{-1} \mathbf{w}_N) + \mathbf{w}_N^T \mathbf{V}_N^{-1} \mathbf{w}_N + \tilde{\mathbf{y}}^T \tilde{\mathbf{y}} + 2b_N \right)^{-\left(\frac{m}{2}+a_N\right)} \\ &\propto \left( 1 + \frac{(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{w}_N)^T \left( \frac{b_N}{a_N} (\mathbf{I}_m + \tilde{\mathbf{X}}\mathbf{V}_N\tilde{\mathbf{X}}^T) \right)^{-1} (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{w}_N)}{2a_N} \right)^{-\left(\frac{m}{2}+a_N\right)} \leftarrow \text{Use the } \text{Sherman Morrison Woodbury formula} \text{ here to show that (symmetry of } \mathbf{V}_0 \text{ is assumed)} \\ &\quad (\mathbf{I}_m + \tilde{\mathbf{X}}\mathbf{V}_N\tilde{\mathbf{X}}^T)^{-1} = \mathbf{I}_m - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} + \mathbf{V}_N^{-1})^{-1} \tilde{\mathbf{X}}^T \end{aligned}$$



# Bayesian Inference when $\sigma^2$ is Unknown

- The posterior predictive is also a Student's  $\mathcal{T}$ :

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathcal{D}) = \mathcal{T}_m \left( \tilde{\mathbf{y}}|\tilde{\mathbf{X}}\mathbf{w}_N, \frac{b_N}{a_N} (\mathbf{I}_m + \tilde{\mathbf{X}}\mathbf{V}_N\tilde{\mathbf{X}}^T), 2a_N \right)$$

- The predictive variance has two terms

- $\frac{b_N}{a_N} \mathbf{I}_m$  due to the measurement noise
- and  $\frac{b_N}{a_N} \tilde{\mathbf{X}}\mathbf{V}_N\tilde{\mathbf{X}}^T$  due to the uncertainty in  $\mathbf{w}$ . *The second term depends on how close a test input is to the training data.*



# Zellner's $G$ -Prior

$$p(\mathbf{w}, \sigma^2) = \mathcal{N}\mathcal{IG}(\mathbf{w}, \sigma^2 | \mathbf{w}_0, \mathbf{V}_0, a_0, b_0) \triangleq \mathcal{N}(\mathbf{w} | \mathbf{w}_0, \sigma^2 \mathbf{V}_0) \mathcal{IG}(\sigma^2 | a_0, b_0)$$

- It is common to set  $a_0 = b_0 = 0$ , corresponding to an uninformative prior for  $\sigma^2$ , and to set  $\mathbf{w}_0 = \mathbf{0}$  and  $\mathbf{V}_0 = g(\mathbf{X}^T \mathbf{X})^{-1}$  for any positive value  $g$ .
- This is called Zellner's **g-prior**. Here  $g$  plays a role analogous to  $1/\lambda$  in ridge regression. However, the prior covariance is proportional to  $(\mathbf{X}^T \mathbf{X})^{-1}$  rather than  $\mathbf{I}$ .

$$p(\mathbf{w}, \sigma^2) = \mathcal{N}\mathcal{IG}(\mathbf{w}, \sigma^2 | \mathbf{0}, g(\mathbf{X}^T \mathbf{X})^{-1}, 0, 0) \triangleq \mathcal{N}(\mathbf{w} | \mathbf{0}, \sigma^2 g(\mathbf{X}^T \mathbf{X})^{-1}) \mathcal{IG}(\sigma^2 | 0, 0)$$

- This ensures that the posterior is invariant to scaling of the inputs.

- Zellner, A. (1986). [On assessing prior distributions and bayesian regression analysis with g-prior distributions](#). In [Bayesian inference and decision techniques](#), Studies of Bayesian and Econometrics and Statistics volume 6. North Holland.
- [Minka, T. \(2000b\). Bayesian linear regression](#). Technical report, MIT.



# Unit Information Prior

$$p(\mathbf{w}, \sigma^2) = \mathcal{N}\mathcal{I}\mathcal{G}(\mathbf{w}, \sigma^2 | 0, g(\mathbf{X}^T \mathbf{X})^{-1}, 0, 0) \triangleq \mathcal{N}(\mathbf{w} | 0, \sigma^2 g(\mathbf{X}^T \mathbf{X})^{-1}) \mathcal{I}\mathcal{G}(\sigma^2 | 0, 0)$$

- ❑ We will see below that if we use an uninformative prior, the posterior precision given  $N$  measurements is  $V_N^{-1} = \mathbf{X}^T \mathbf{X}$ .
- ❑ The **unit information prior** is defined to contain as much information as one sample.
- ❑ *To create a unit information prior for linear regression, we need to use  $V_0^{-1} = \frac{1}{N} \mathbf{X}^T \mathbf{X}$  which is equivalent to the g-prior with  $g = N$ .*
- ❑ *Zellner's prior depends on the data: This is contrary to much of our Bayesian inference discussion!*

- [Kass, R. and L. Wasserman \(1995\)](#). A reference bayesian test for nested [hypotheses and its relationship to the schwarz criterion](#). *J. of the Am. Stat. Assoc.* 90(431), 928–934.



# Uninformative Prior

- An uninformative prior can be obtained by considering the *uninformative limit of the conjugate g-prior, which corresponds to setting  $g = \infty$* . This is equivalent to an improper  $\mathcal{NIG}$  prior with  $\mathbf{w}_0 = 0$ ,  $\mathbf{V}_0 = \infty \mathbf{I}$ ,  $a_0 = 0$  and  $b_0 = 0$ , which gives  $p(\mathbf{w}, \sigma^2) \propto \sigma^{-(D+2)}$ .

$$p(\mathbf{w}, \sigma^2) = \mathcal{NIG}(\mathbf{w}, \sigma^2 | 0, \infty \mathbf{I}, 0, 0) \triangleq \mathcal{N}(\mathbf{w} | 0, \sigma^2 \infty \mathbf{I}) \mathcal{IG}(\sigma^2 | 0, 0) \rightarrow \sigma^{-(D+2)}$$

- Alternatively, we can start with the *semi-conjugate prior  $p(\mathbf{w}, \sigma^2) = p(\mathbf{w})p(\sigma^2)$ , and take each term to its uninformative limit individually*, which gives  $p(\mathbf{w}, \sigma^2) \propto \sigma^{-2}$ .

This is equivalent to an improper  $\mathcal{NIG}$  prior with  $\mathbf{w}_0 = \mathbf{0}$ ,  $\mathbf{V} = \infty \mathbf{I}$ ,  $a_0 = -D/2$  and  $b_0 = 0$ .

$$p(\mathbf{w}, \sigma^2) = \mathcal{NIG}(\mathbf{w}, \sigma^2 | 0, \infty \mathbf{I}, 0, 0) \triangleq \mathcal{N}(\mathbf{w} | 0, \sigma^2 \infty \mathbf{I}) \mathcal{IG}\left(\sigma^2 | -\frac{D}{2}, 0\right) \rightarrow \sigma^{-2}$$



# Uninformative Prior

- Using the uninformative prior,  $p(\mathbf{w}, \sigma^2) \propto \sigma^{-2}$ , the corresponding posterior and marginal posteriors are given by

$$p(\mathbf{w}, \sigma^2 | \mathcal{D}) = \mathcal{NIG}(\mathbf{w}, \sigma^2 | \mathbf{w}_N, \mathbf{V}_N, a_N, b_N)$$

$$p(\mathbf{w} | \mathcal{D}) = \mathcal{I}_D \left( \mathbf{w}_N, \frac{b_N}{a_N} \mathbf{V}_N, 2a_N \right) = \mathcal{I}_D \left( \mathbf{w} | \hat{\mathbf{w}}_{MLE}, \frac{s^2}{N-D} \mathbf{C}, N-D \right)$$

$$\mathbf{V}_N = \mathbf{C} = (\mathbf{V}_0^{-1} + \mathbf{X}^T \mathbf{X})^{-1} \rightarrow (\mathbf{X}^T \mathbf{X})^{-1}, \quad \mathbf{w}_N = \mathbf{V}_N (\mathbf{V}_0^{-1} \mathbf{w}_0 + \mathbf{X}^T \mathbf{y}) \rightarrow (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \hat{\mathbf{w}}_{MLE}$$

$$a_N = a_0 + N/2 = (N-D)/2,$$

$$b_N = b_0 + \frac{1}{2} (\mathbf{w}_0^T \mathbf{V}_0^{-1} \mathbf{w}_0 + \mathbf{y}^T \mathbf{y} - \mathbf{w}_N^T \mathbf{V}_N^{-1} \mathbf{w}_N) = s^2/2, \quad s^2 = (\mathbf{y} - \mathbf{X} \hat{\mathbf{w}}_{MLE})^T (\mathbf{y} - \mathbf{X} \hat{\mathbf{w}}_{MLE})$$

$$\mathbf{w}_N = \hat{\mathbf{w}}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- Note in the calculation of  $s^2$ :

$$\begin{aligned} s^2 &= (\mathbf{y} - \mathbf{X} \hat{\mathbf{w}}_{MLE})^T (\mathbf{y} - \mathbf{X} \hat{\mathbf{w}}_{MLE}) = (\mathbf{y} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y})^T (\mathbf{y} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}) \\ &= \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{y}^T \mathbf{y} - \hat{\mathbf{w}}_{MLE}^T (\mathbf{X}^T \mathbf{X}) (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{y}^T \mathbf{y} - \hat{\mathbf{w}}_{MLE}^T \mathbf{V}_N^{-1} \hat{\mathbf{w}}_{MLE} \end{aligned}$$



# Frequentist Confidence Interval Vs. Bayesian Marginal Crjedible Interval

---

- The use of a (semi-conjugate) uninformative prior is quite interesting since the resulting posterior *turns out to be equivalent to the results obtained from frequentist statistics.*

$$p(\mathbf{w}_j|D) = \mathcal{T}\left(w_j|\hat{w}_j, \frac{C_{jj}s^2}{N-D}, N-D\right)$$

- This is equivalent to *the sampling distribution of the MLE* which is given by the following:

$$\frac{w_j - \hat{w}_j}{s_j} \sim \mathcal{T}_{N-D}, \quad s_j = \sqrt{\frac{C_{jj}s^2}{N-D}}$$

$s_j$  is the standard error of the estimated parameter.

- *The frequentist confidence interval and the Bayesian marginal credible interval for the parameters are the same.*

- Rice, J. (1995). [Mathematical statistics and data analysis](#). Duxbury. 2<sup>nd</sup> edition (page 542)
- Casella, G. and R. Berger (2002). [Statistical inference](#). Duxbury. 2<sup>nd</sup> edition (page 554)



# The Caterpillar Example

□ As a worked example of the uninformative prior, consider the [caterpillar dataset](#). We can compute the posterior mean and standard deviation, and the 95% credible intervals (CI) for the regression coefficients.

□ The 95% credible intervals are identical to the 95% confidence intervals computed using standard frequentist methods.	coeff	mean	stddev	95pc CI	sig
	w0	10.998	3.06027	[ 4.652, 17.345]	*
	w1	-0.004	0.00156	[ -0.008, -0.001]	*
	w2	-0.054	0.02190	[ -0.099, -0.008]	*
	w3	0.068	0.09947	[ -0.138, 0.274]	
	w4	-1.294	0.56381	[ -2.463, -0.124]	*
	w5	0.232	0.10438	[ 0.015, 0.448]	*
	w6	-0.357	1.56646	[ -3.605, 2.892]	
	w7	-0.237	1.00601	[ -2.324, 1.849]	
	w8	0.181	0.23672	[ -0.310, 0.672]	
	w9	-1.285	0.86485	[ -3.079, 0.508]	
	w10	-0.433	0.73487	[ -1.957, 1.091]	

Run [linregBayesCaterpillar](#)  
from PMTK3

■ Marin, J.-M. and C. Robert (2007). [Bayesian Core: a practical approach to computational Bayesian statistics](#). Springer.



# The Caterpillar Example

---

- ❑ We can use these marginal posteriors to compute if the coefficients are significantly different from 0 -- check if its 95% CI excludes 0.
  - ❑ The CIs for coefficients 0, 1, 2, 4, 5 are all significant.
  - ❑ These results are the same as those produced by a frequentist approach using p-values at the 5% level.
  - ❑ But note that the MLE does not even exist when  $N < D$ , so standard frequentist inference theory breaks down in this setting. Bayesian inference theory still works using proper priors.
- [Maruyama, Y. and E. George](#) (2008). [A g-prior extension for  \$p > n\$](#) . Technical report, U. Tokyo.



# Empirical Bayes for Linear Regression

- We describe next an empirical Bayes procedure for picking the hyper-parameters in the prior.
- More precisely, we choose  $\boldsymbol{\eta} = (\alpha, \lambda)$  to maximize the marginal likelihood, where  $\lambda = 1/\sigma^2$  be the precision of the observation noise and  $\alpha$  is the precision of the prior,  $p(\mathbf{w}) = N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$ .
- This is known as the *evidence procedure*.

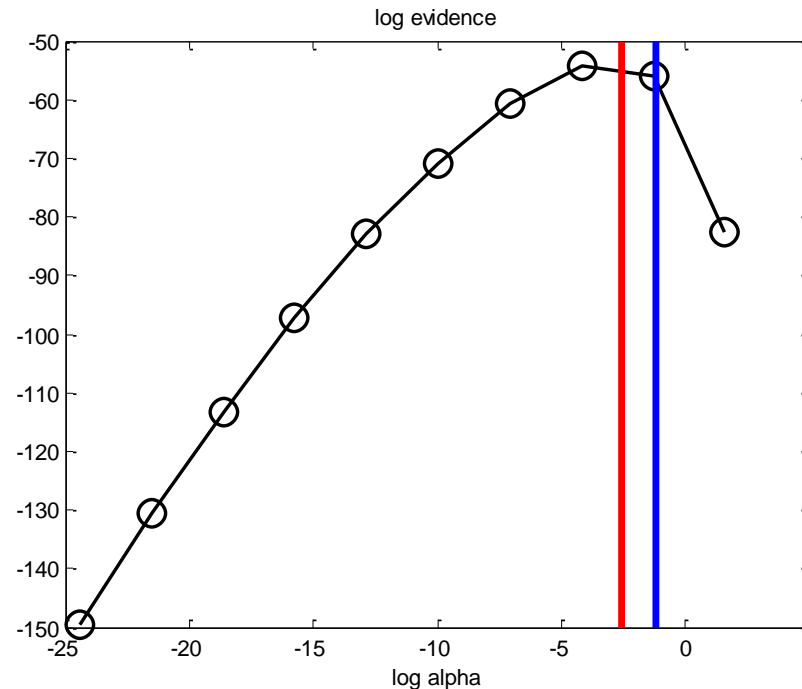
- [MacKay, D. \(1995b\). Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. \*Network\*.](#)
- [Buntine, W. and A. Weigend \(1991\). Bayesian backpropagation. \*Complex Systems\* 5, 603–643.](#)
- [MacKay, D. \(1999\). Comparision of approximate methods for handling hyperparameters. \*Neural Computation\* 11\(5\), 1035–1068.](#)





# Empirical Bayes for Linear Regression

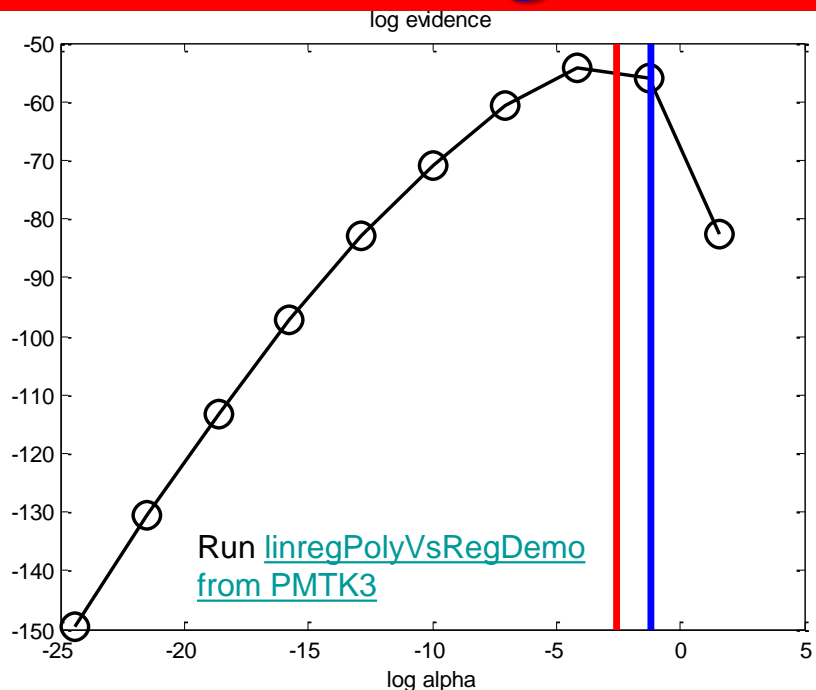
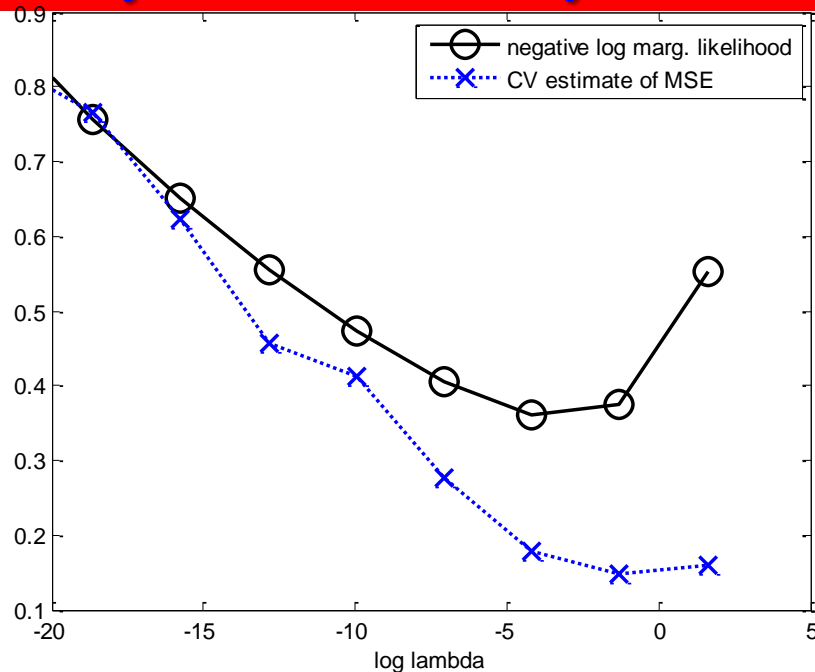
- ❑ The evidence procedure provides an alternative to using cross validation.
- ❑ In the Figure, the log marginal likelihood is plotted for different values of  $\alpha$ , as well as the maximum value found by the optimizer.



Run [linregPolyVsRegDemo](#)  
from PMTK3



# Empirical Bayes for Linear Regression



- We obtain the same result as 5-CV ( $\lambda = 1/\sigma^2$  is fixed in both methods).
- The key advantage of the evidence procedure over CV is that it allows different  $\alpha_j$  to be used for every feature.

# The Evidence Approximation

---

- The evidence procedure can be used to perform feature selection.
- The evidence procedure is also useful when comparing different kinds of models:

$$\begin{aligned} p(\mathcal{D} | m) &= \iint p(\mathcal{D} | \mathbf{w}, m) p(\mathbf{w} | m, \boldsymbol{\eta}) p(\boldsymbol{\eta} | m) d\mathbf{w} d\boldsymbol{\eta} \\ &\approx \max_{\boldsymbol{\eta}} \int p(\mathcal{D} | \mathbf{w}, m) p(\mathbf{w} | m, \boldsymbol{\eta}) p(\boldsymbol{\eta} | m) d\mathbf{w} \end{aligned}$$

- It is important to (at least approximately) integrate over  $\boldsymbol{\eta}$  rather than setting it arbitrarily.
- Using variation Bayes models our uncertainty on  $\boldsymbol{\eta}$  rather than computing point estimates.



# Bayesian Model Comparison

- The Bayesian view of model comparison involves the use of probabilities to represent uncertainty in the choice of model.
- Suppose we wish to compare  $L$  models  $\{\mathcal{M}_i\}$  where  $i = 1, \dots, L$ . Here a model refers to a probability distribution over the observed data  $\mathcal{D}$ . Our uncertainty is expressed through a prior  $p(\mathcal{M}_i)$ . Given a training set  $\mathcal{D}$ , then the posterior distribution is:

$$p(\mathcal{M}_i | \mathcal{D}) \propto p(\mathcal{M}_i) \underbrace{p(\mathcal{D} | \mathcal{M}_i)}$$

Posterior

Prior

**Model evidence or  
marginal likelihood**

- We have already defined the **Bayes Factor** as the ratio of two model evidences

$$\frac{p(\mathcal{D} | \mathcal{M}_i)}{p(\mathcal{D} | \mathcal{M}_j)}$$



# Model Averaging and Model Selection

- Once we know the posterior distribution over models, the **predictive distribution** is given, by


$$p(t \mid \underset{\substack{\uparrow \\ \text{Test data}}}{\mathbf{x}}, \mathcal{D}) = \sum_{i=1}^L p(t \mid \mathbf{x}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i \mid \underset{\substack{\uparrow \\ \text{Training data}}}{\mathcal{D}})$$

- This has the form of a **mixture distribution** in which the overall predictive distribution is obtained by **averaging the predictive distributions**  $p(t \mid \mathbf{x}, \mathcal{M}_i, \mathcal{D})$  **of individual models,** weighted by the posterior probabilities  $p(\mathcal{M}_i \mid \mathcal{D})$  **of those models.**
- A simple approximation to model averaging is to use the single most probable model alone to make predictions. This is known as **model selection**.




# Model Evidence

- For a model governed by a set of parameters  $\mathbf{w}$ , the model evidence is given, from the sum and product rules of probability, by

$$p(\mathcal{D} | \mathcal{M}_i) = \int p(\mathcal{D} | \mathbf{w}_i, \mathcal{M}_i) p(\mathbf{w}_i | \mathcal{M}_i) d\mathbf{w}_i$$


- From a sampling perspective, **the marginal likelihood can be viewed as the probability of generating the data set  $\mathcal{D}$  from a model whose parameters are sampled at random from the prior.**

- Also we can see the model evidence as normalizing factor:

$$p(\mathbf{w} | \mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D} | \mathbf{w}, \mathcal{M}_i) p(\mathbf{w} | \mathcal{M}_i)}{p(\mathcal{D} | \mathcal{M}_i)}$$


# Occam's Razor and Model Selection

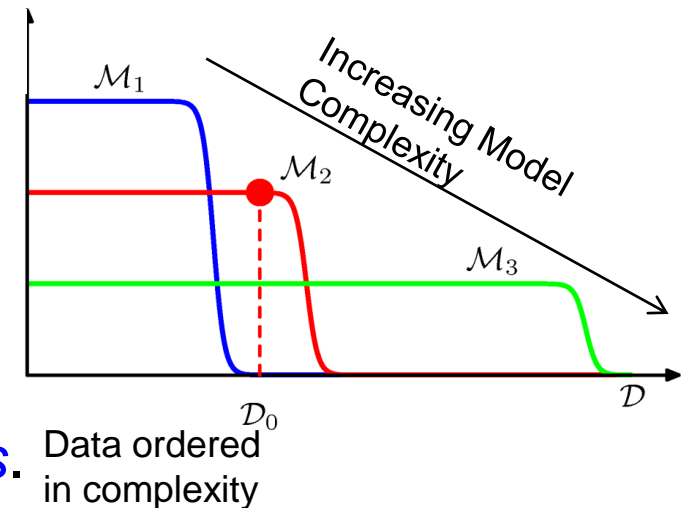
- Compare model classes  $\mathcal{M}_i$  using their posterior probability given the data  $\mathcal{D}$  :

$$p(\mathcal{M}_i | \mathcal{D}) \propto p(\mathcal{M}_i) p(\mathcal{D} | \mathcal{M}_i) \quad p(\mathcal{D} | \mathcal{M}_i) = \int_{\Theta_i} p(\mathcal{D} | \mathbf{w}_i, \mathcal{M}_i) p(\mathbf{w}_i | \mathcal{M}_i) d\mathbf{w}_i$$

- The marginal likelihood (Bayesian evidence)  $p(\mathcal{D} | \mathcal{M}_i)$  is viewed as **the probability that randomly selected parameter values from the model class would generate the data set  $\mathcal{D}$ .**

- Simple model classes are unlikely to generate  $\mathcal{D}$ .
- Too complex model classes  $p(\mathcal{D} | \mathcal{M}_i)$  can generate many data sets so it is unlikely to generate the particular data set  $\mathcal{D}$ .

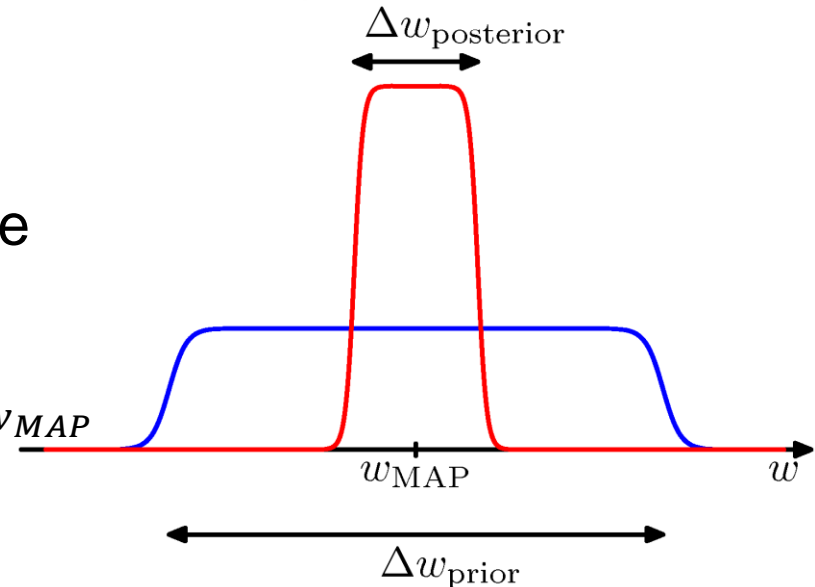
- Bayesian inference automatically implements Occam's Razor Principle:  
*Prefer simple than complex explanations.*



# Bayesian Model Comparison

- For a given model (omit  $i$  |  $\mathcal{M}_i$ ) with a single parameter,  $w$ , consider the approximation (use Bayes rule)

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)dw = \frac{p(\mathcal{D}|w)p(w)}{p(w|\mathcal{D})} \Big|_{w_{MAP}} \\ \simeq p(\mathcal{D}|w_{MAP}) \frac{\Delta w_{posterior}}{\Delta w_{prior}}$$



where the posterior is assumed to be sharply peaked around  $w_{MAP}$  (we already have seen the more accurate Laplace approximation)

- Taking logs we obtain

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{MAP}) + \underbrace{\ln \left( \frac{\Delta w_{posterior}}{\Delta w_{prior}} \right)}_{\text{Negative}}$$

Note: the evidence is not defined if the prior is improper.





# Optimal Model Complexity

- For a model with  $M$  parameters, we can make a similar approximation for each parameter in turn. Assuming that all parameters have the same ratio of  $\Delta w_{\text{posterior}} / \Delta w_{\text{prior}}$ ,

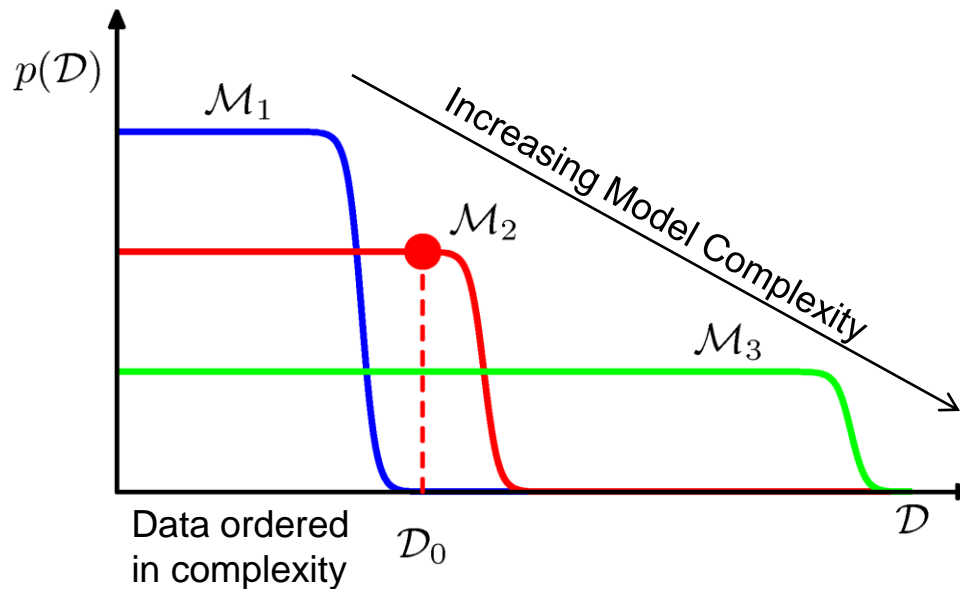
$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D} | w_{MAP}) + M \ln \left( \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)$$

- The size of the complexity penalty increases linearly with  $M$ . As we increase the complexity of the model
  - the 1<sup>st</sup> term increases, because a more complex model is better able to fit the data,
  - whereas the 2<sup>nd</sup> term decreases due to the dependence on  $M$ .
- The optimal model complexity determined by maximum evidence will be given by a trade-off between these two terms.



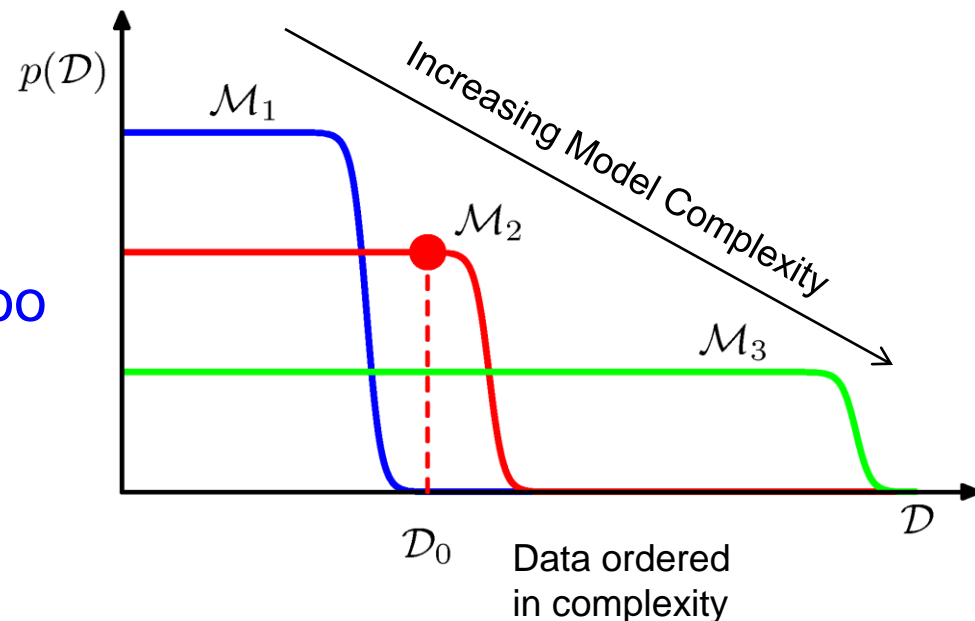
# Matching Data and Model Complexity

- The marginal likelihood favors models of intermediate complexity.
- Let us think of the regression model and consider the models  $\mathcal{M}_1, \mathcal{M}_2$  and  $\mathcal{M}_3$  represent linear, quadratic and cubic fitting.
- The data  $\mathcal{D}$  are ordered in complexity – for a given model, we choose  $\mathbf{w}$  from the prior  $p(\mathbf{w})$ , then sample the data from  $p(\mathcal{D}|\mathbf{w})$ .



# Matching Data and Model Complexity

- A 1<sup>st</sup> order polynomial has little variability, generates data that are similar,  $p(\mathcal{D})$  is confined to a small region in the  $\mathcal{D}$  axis.
- A 9<sup>th</sup> order polynomial generates a variety of different data, and so its  $p(\mathcal{D})$  is spread over a large region in the  $\mathcal{D}$  axis.
- Because  $p(\mathcal{D}|M_i)$  are normalized, a particular  $\mathcal{D}_0$  can have the highest evidence for the model of intermediate complexity.
- The simpler model cannot fit the data well, whereas the more complex model spreads its predictive probability over too broad a range of data sets.



# Matching Data and Model Complexity

- A Bayesian model comparison in an average (over the data  $\mathcal{D}$ ) sense will favor the correct model.
- Let  $\mathcal{M}_1$  be the correct model and  $\mathcal{M}_2$  another model. We can show that the evidence for model  $\mathcal{M}_1$  is higher. Using the definition and properties of the [Kullback-Leibler](#) distance:

$$KL(p(\mathcal{D} | \mathcal{M}_1) \| p(\mathcal{D} | \mathcal{M}_2)) = \int \underbrace{p(\mathcal{D} | \mathcal{M}_1)}_{\text{Averaged with the exact probability}} \ln \underbrace{\frac{p(\mathcal{D} | \mathcal{M}_1)}{p(\mathcal{D} | \mathcal{M}_2)}}_{\text{Bayes factor}} d\mathcal{D} \geq 0$$

- This analysis assumes that the true distribution from which the data are generated is contained in our class of models.



# The Evidence Approximation

- The fully Bayesian predictive distribution for our regression model is given by

The hyperparameters  $\alpha$  and  $\beta$  are now random variables

$$p(t|\mathbf{t}) = \int \int \int \underbrace{p(t|\mathbf{w}, \beta)}_{\mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})} \underbrace{p(\mathbf{w}|\mathbf{t}, \alpha, \beta)}_{\substack{\mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \\ \mathbf{m}_N = \beta \mathbf{S}_N \Phi^T \mathbf{t} \\ \mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \Phi^T \Phi}} p(\alpha, \beta|\mathbf{t}) d\mathbf{w} d\alpha d\beta$$

Dependence on  $x$  and  $\mathbf{x}$  not shown to simplify the notation

but this integral is intractable. Approximate with

$$p(t|\mathbf{t}) \simeq p(t|\mathbf{t}, \hat{\alpha}, \hat{\beta}) = \int p(t|\mathbf{w}, \hat{\beta}) p(\mathbf{w}|\mathbf{t}, \hat{\alpha}, \hat{\beta}) d\mathbf{w}$$

where  $(\hat{\alpha}, \hat{\beta})$  is the mode of  $p(\alpha, \beta|\mathbf{t})$ , which is assumed to be sharply peaked.

a.k.a. empirical Bayes, type II or generalized maximum likelihood, or evidence approximation.



# The Evidence Approximation

- From Bayes' theorem, the posterior distribution for  $\alpha$  and  $\beta$  is given by

$$p(\alpha, \beta | \mathbf{t}) \propto p(\mathbf{t} | \alpha, \beta) p(\alpha, \beta)$$

- If the prior is relatively flat, then in the evidence framework the values of  $(\hat{\alpha}, \hat{\beta})$  are obtained by maximizing the marginal likelihood function  $p(\mathbf{t} | \alpha, \beta)$ .

- The marginal likelihood function  $p(\mathbf{t} | \alpha, \beta)$  is obtained by integrating over the parameters  $\mathbf{w}$ , so that

$$p(\alpha, \beta | \mathbf{t}) \propto p(\mathbf{t} | \alpha, \beta) = \int \underbrace{p(\mathbf{t} | \mathbf{w}, \beta)}_{\text{Evidence/}} \underbrace{p(\mathbf{w} | \alpha)}_{\text{Marginal Likelihood}} d\mathbf{w}$$

$\frac{1}{(2\pi)^{N/2}} \beta^{N/2} e^{-\beta E_D} \mathcal{N}(\mathbf{w}, \alpha^{-1} \mathbf{I}_{M \times M})$

- One can evaluate this integral using the completion of the square procedure typical of Gaussian marginalizations.

# The Evidence Approximation

- We can write the evidence function in the form

$$p(\mathbf{t} / \alpha, \beta) = \left( \frac{\beta}{2\pi} \right)^{N/2} \left( \frac{\alpha}{2\pi} \right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

where  $M$  is the dimensionality of  $\mathbf{w}$ , and we have defined

$$E(\mathbf{w}) = \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$$
$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^T \mathbf{A} (\mathbf{w} - \mathbf{m}_N)$$

- We have introduced here:

$$\mathbf{A} = \alpha \mathbf{I} + \beta \Phi^T \Phi, \quad E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^T \mathbf{m}_N, \quad \mathbf{m}_N = \beta \mathbf{A}^{-1} \Phi^T \mathbf{t}$$

- Note that the Hessian matrix  $\mathbf{A}$  corresponds to the matrix of 2<sup>nd</sup> derivatives of the error function:

$$\mathbf{A} = \nabla \nabla E(\mathbf{w})$$



# The Evidence Approximation

- The integral over  $\mathbf{w}$  can now be evaluated simply by appealing to the standard result for the normalization coefficient of a multivariate Gaussian, giving

$$\begin{aligned}\int \exp\{-E(\mathbf{w})\} d\mathbf{w} &= \exp\{-E(\mathbf{m}_N)\} \int \exp\left\{-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T \mathbf{A}(\mathbf{w} - \mathbf{m}_N)\right\} d\mathbf{w} \\ &= \exp\{-E(\mathbf{m}_N)\} (2\pi)^{M/2} |\mathbf{A}|^{-1/2}\end{aligned}$$

- We can then write the log of the marginal likelihood in the form

$$p(\mathbf{t} | \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w} = \left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} e^{-E(\mathbf{m}_N)} (2\pi)^{M/2} |\mathbf{A}|^{-1/2} \Rightarrow$$

$$\ln p(\mathbf{t} | \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\mathbf{m}_N) - \frac{1}{2} \ln |\mathbf{A}| - \frac{N}{2} \ln (2\pi)$$

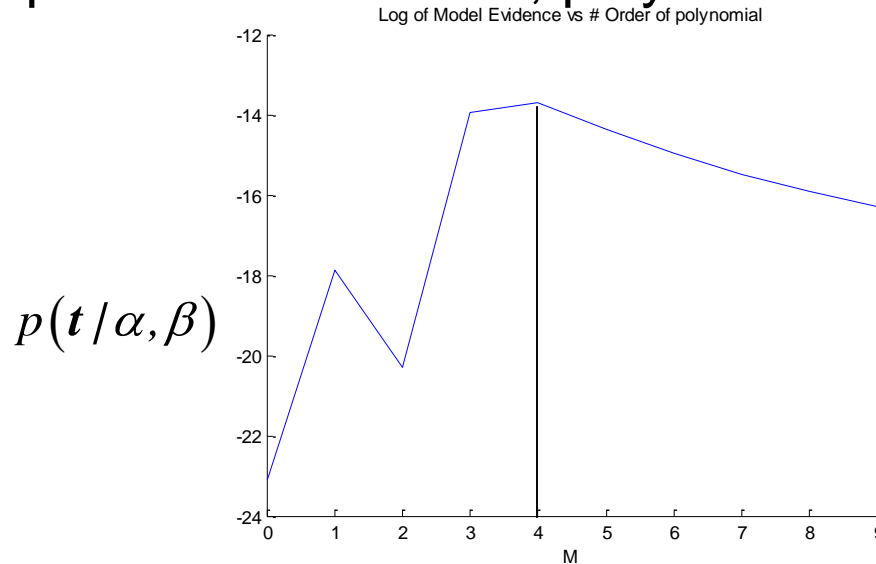
*M : number of parameters in the model*

*N : number of data*



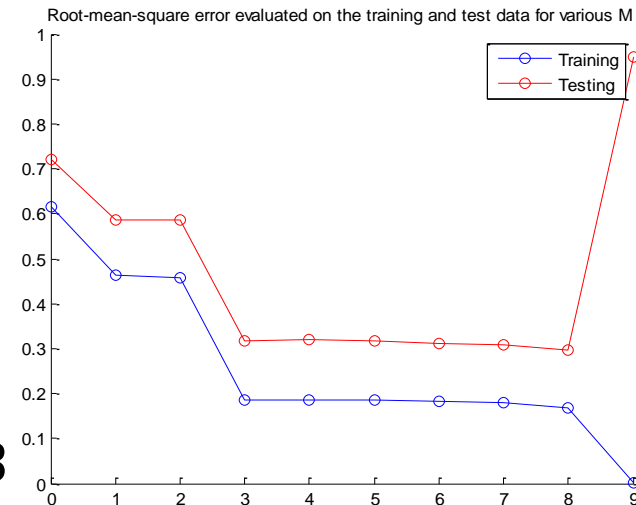
# The Evidence Approximation

- Example: sinusoidal data, polynomial regression,  $\alpha = 5 \times 10^{-3}$ ,  
 $\beta = 11.1$



[MatLab Code](#)  
and [data](#)

- From the plot of the model evidence for given  $\alpha$  and  $\beta$ , we see that the evidence favors the model with  $M = 5$  (4<sup>th</sup> degree polynomial)
- Looking at the non-Bayesian approach, one cannot distinguish the performance of polynomials of order 3 ... 8



# Maximizing the Evidence Function

- Let us first consider the maximization of  $p(\mathbf{t}|\alpha, \beta)$  with respect to  $\alpha$ . This can be done by first defining the following eigenvector equation

$$(\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

- Thus,  $\mathbf{A} = \alpha \mathbf{I} + \beta \Phi^T \Phi$  has eigenvalues  $\alpha + \lambda_i$ .
- Now consider the derivative of the term involving  $\ln |\mathbf{A}|$  with respect to  $\alpha$ . We have

$$\ln p(\mathbf{t} | \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\mathbf{m}_N) - \frac{1}{2} \ln |\mathbf{A}| - \frac{N}{2} \ln(2\pi)$$

$$\frac{d}{d\alpha} \ln |\mathbf{A}| = \frac{d}{d\alpha} \ln \prod_i (\lambda_i + \alpha) = \frac{d}{d\alpha} \sum_i \ln(\lambda_i + \alpha) = \sum_i \frac{1}{\lambda_i + \alpha}$$

# Maximizing the Evidence Function

$$\frac{d}{d\alpha} \ln |A| = \frac{d}{d\alpha} \ln \prod_i (\lambda_i + \alpha) = \frac{d}{d\alpha} \sum_i \ln(\lambda_i + \alpha) = \sum_i \frac{1}{\lambda_i + \alpha}$$

□ Thus the stationary points of

$$\ln p(\mathbf{t} | \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\mathbf{m}_N) - \frac{1}{2} \ln |A| - \frac{N}{2} \ln(2\pi)$$

$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^T \mathbf{m}_N, \quad \mathbf{m}_N = \beta A^{-1} \Phi^T \mathbf{t}, \quad A = \alpha \mathbf{I} + \beta \Phi^T \Phi$$

with respect to  $\alpha$  satisfy 
$$0 = \frac{M}{2\alpha} - \frac{1}{2} \mathbf{m}_N^T \mathbf{m}_N - \frac{1}{2} \sum_i \frac{1}{\lambda_i + \alpha}$$

□ Multiplying through by  $2\alpha$  and rearranging, we obtain

$$\alpha \mathbf{m}_N^T \mathbf{m}_N = M - \alpha \sum_i \frac{1}{\lambda_i + \alpha} = \sum_i \left( 1 - \frac{\alpha}{\lambda_i + \alpha} \right) = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} \equiv \gamma$$

$$\alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N}$$

Implicit solution  
for  $a$

1. Choose  $a$
2. Calculate  $\mathbf{m}_N, \gamma$
3. Re-estimate  $a$



# Maximizing the Evidence Function

## Implicit Solution for Computing $\alpha$

1. Choose  $\alpha$
2. Calculate

$$\mathbf{m}_N, \gamma:$$

$$\mathbf{m}_N = \beta \mathbf{A}^{-1} \Phi^T \mathbf{t}, \quad \mathbf{A} = a \mathbf{I} + \beta \Phi^T \Phi, \quad (\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

3. Re-estimate  $a$

$$\gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha}$$

$$\alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N}$$

# Maximizing the Evidence Function

- We can similarly maximize the log marginal likelihood with respect to  $\beta$ .

$$\ln p(\mathbf{t} \mid \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\mathbf{m}_N) - \frac{1}{2} \ln | \mathbf{A} | - \frac{N}{2} \ln (2\pi)$$

- To do this, we note that the eigenvalues  $\lambda_i$  defined by

$$(\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

are proportional to  $\beta$ , and hence  $d\lambda_i / d\beta = \lambda_i / \beta$  giving

$$\frac{d}{d\beta} \ln | \mathbf{A} | = \frac{d}{d\beta} \ln \prod_i (\lambda_i + \alpha) = \frac{1}{\beta} \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}$$

# Maximizing the Evidence Function

$$\ln p(\mathbf{t} | \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\mathbf{m}_N) - \frac{1}{2} \ln |A| - \frac{N}{2} \ln(2\pi)$$

$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^T \mathbf{m}_N$$

$$\frac{d}{d\beta} \ln |A| = \frac{1}{\beta} \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}$$

$$\mathbf{m}_N = \beta A^{-1} \Phi^T \mathbf{t}, \quad A = \alpha \mathbf{I} + \beta \Phi^T \Phi, \quad (\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad \gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha}$$

□ Setting the derivative wrt  $\beta$  equal to zero, the stationary point of the marginal likelihood therefore satisfies

$$0 = \frac{N}{2\beta} - \frac{1}{2} \sum_{n=1}^N \left\{ t_n - \mathbf{m}_N^T \boldsymbol{\phi}(x_n) \right\}^2 - \frac{\gamma}{2\beta}$$

□ Rearranging we obtain

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^N \left\{ t_n - \mathbf{m}_N^T \boldsymbol{\phi}(x_n) \right\}^2$$

Implicit solution  
for  $\beta$

1. Choose  $\beta$
2. Calculate  $\mathbf{m}_N, \gamma$
3. Re-estimate  $\beta$



# Maximizing the Evidence Function

- It is interesting to note that in the evidence framework (using the optimal computed values of  $\alpha$  and  $\beta$ ), the following is true:

$$E(\mathbf{m}_N) = \frac{N}{2}$$

- This can be easily shown using the earliest derived results:

$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^T \mathbf{m}_N$$

with

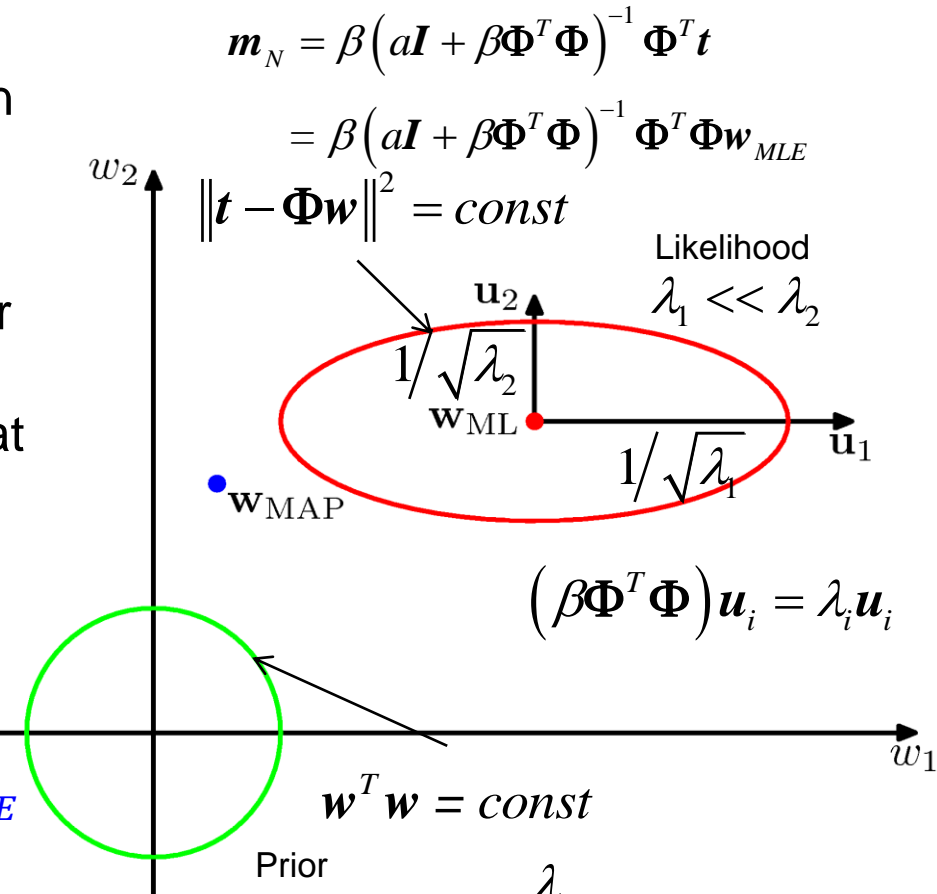
$$\alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N}$$

and

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^N \left\{ t_n - \mathbf{m}_N^T \boldsymbol{\phi}(x_n) \right\}^2 = \frac{1}{N - \gamma} \|\mathbf{t} - \Phi \mathbf{m}_N\|^2$$

# Effective Number of Parameters

- Consider the contours of the likelihood & prior in which the axes in parameter space have been rotated to align with the eigenvectors  $\mathbf{u}_i$ .
- For  $\alpha = 0$ , the mode of the posterior is given by the MLE solution  $\mathbf{w}_{ML}$ , whereas for nonzero  $\alpha$  the mode is at  $\mathbf{w}_{MAP} = \mathbf{m}_N$ .
- In the direction  $w_1$ ,  $\lambda_1$  is small compared with  $\alpha$  and  $\lambda_1/(\lambda_1 + \alpha)$  is close to zero, and the corresponding MAP value of  $w_1$ ,  $w_{1MAP} = \frac{\lambda_1}{\lambda_1 + \alpha} w_{1MLE}$  is also close to zero.
- In the direction  $w_2$ ,  $\lambda_2$  is large compared with  $\alpha$  and so  $\lambda_2/(\lambda_2 + \alpha)$  is close to unity, and the MAP value of  $w_2$  is close to its MLE value.



$$0 < \frac{\lambda_i}{\lambda_i + \alpha} \leq 1,$$

$$0 \leq \gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} \leq M$$





# Effective Number of Parameters

- In directions  $w_i, \lambda_i \ll \alpha$ ,  $\lambda_i/(\lambda_i + \alpha)$  is close to zero, and the corresponding MAP value of  $w_i$  is also close to zero.

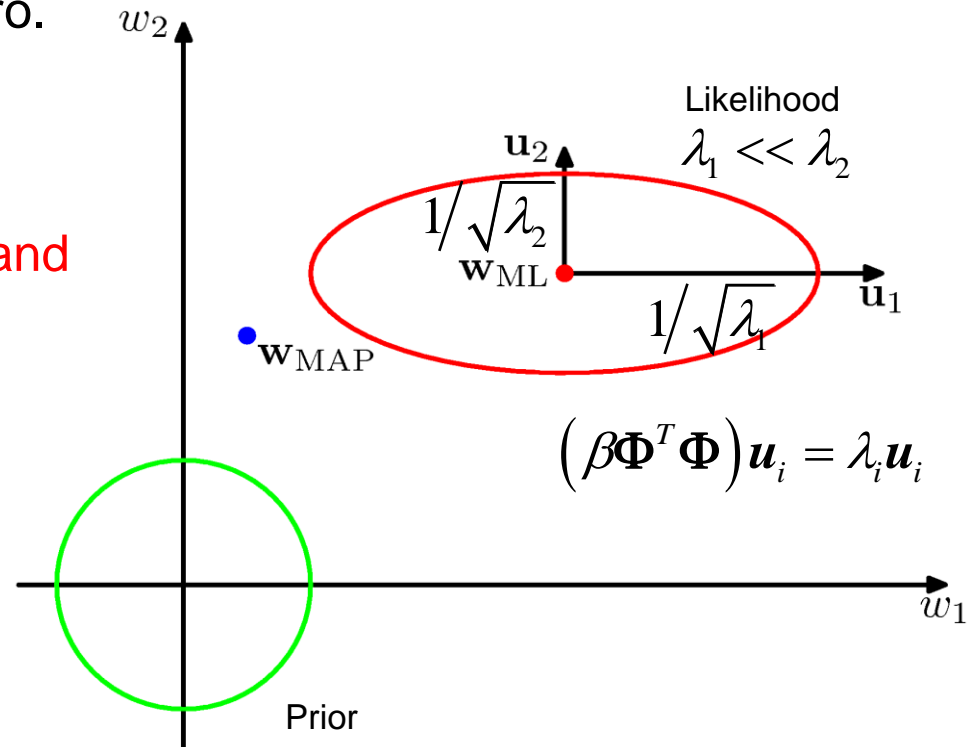
These are directions in which the likelihood function is relatively insensitive to the parameter value and so the parameter has been set to a small value by the prior.

- The quantity  $\gamma$

$$0 < \frac{\lambda_i}{\lambda_i + \alpha} \leq 1,$$

$$0 \leq \gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} \leq M$$

therefore measures the effective total number of well determined parameters.



# Effective Number of Parameters

- We can obtain some insight into the equation for  $\beta$

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^N \left\{ t_n - \mathbf{m}_N^T \boldsymbol{\phi}(x_n) \right\}^2$$

by comparing it with the MLE result [derived earlier](#):

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^N \left\{ t_n - \mathbf{m}_N^T \boldsymbol{\phi}(x_n) \right\}^2$$

- These formulas express the variance as an average of the squared differences between the targets and model predictions.
- They differ in that the # of data points
  - $N$  in the MLE result is replaced by
  - $N - \gamma$  in the Bayesian result.



# Effective Number of Parameters

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^N \left\{ t_n - \mathbf{m}_N^T \boldsymbol{\phi}(x_n) \right\}^2$$

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^N \left\{ t_n - \mathbf{m}_N^T \boldsymbol{\phi}(x_n) \right\}^2$$

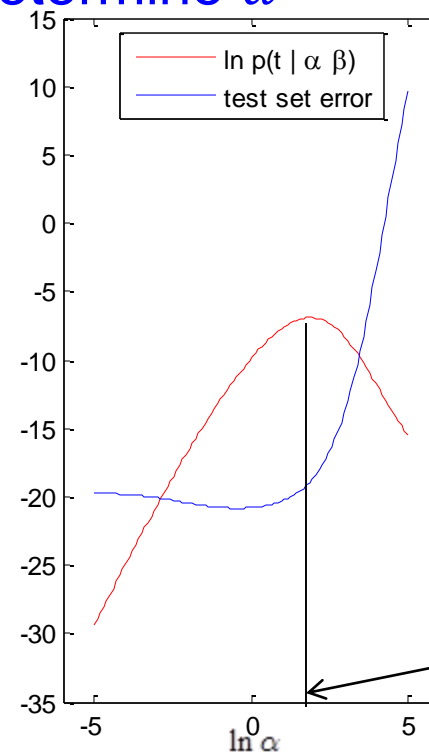
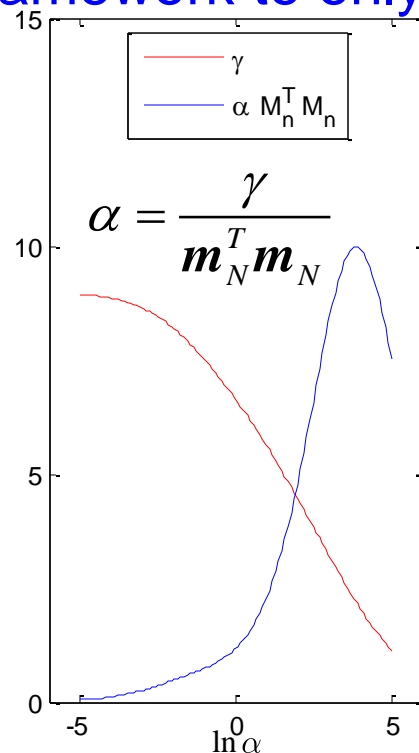
- The effective number of parameters determined by the data is  $\gamma$ .
- The remaining  $M - \gamma$  parameters are set to small values by the prior.
- This is reflected in the Bayesian result for the variance that has a factor  $N - \gamma$  in the denominator correcting for the bias of the MLE.
- These results are analogous to the estimation of the variance of a Gaussian:

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N \left\{ x_n - \mu_{ML} \right\}^2 \quad \text{vs.} \quad \sigma_{MAP}^2 = \frac{1}{N - 1} \sum_{n=1}^N \left\{ x_n - \mu_{ML} \right\}^2$$

1 degree of freedom has been used to fit the mean and the MAP estimate for the variance accounts for that.

# Effective Number of Parameters

- We illustrate the evidence framework for setting hyperparameters using the sinusoidal synthetic data, together with 9 Gaussian basis functions. The total # of parameters is thus  $M = 10$  including the bias.
- For simplicity, we set  $\beta = 11.1$  (true value) and use the evidence framework to only determine  $\alpha$



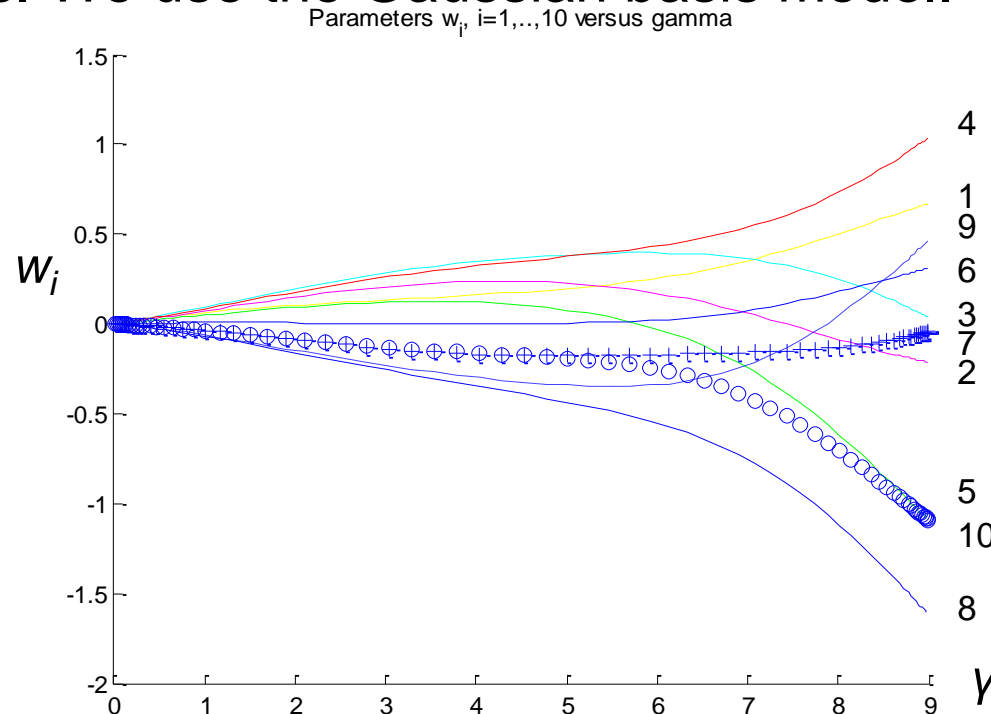
[MatLab Code and data](#)

Min generalization error



# Effective Number of Parameters

- We can also see how  $\alpha$  controls the magnitude of the parameters  $\{w_i\}$ , by plotting the individual parameters (posterior means) versus the effective number  $\gamma$  of parameters. We use the Gaussian basis model.



[MatLab Code](#)  
and [data](#)

- For the simulation,  $\alpha$  is varied  $0 \leq \alpha \leq \infty$  causing  $\gamma$  to vary in the range  $0 \leq \gamma \leq M$ .

# Case of $N \gg M$

- For  $N \gg M$ , all of the parameters are well determined by the data because  $\Phi^T \Phi$  involves an implicit sum over data points, and so the eigenvalues  $\lambda_i$  increase with the size of the data set.
- In this case,  $\gamma = M$ , and the re-estimation equations for  $\alpha$  and  $\beta$  become

$$\alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N} = \frac{M}{2E_W(\mathbf{m}_N)} = \frac{M}{\mathbf{m}_N^T \mathbf{m}_N} \quad (\gamma = M)$$

$$\frac{1}{\beta} = \frac{1}{N} \sum_{n=1}^N \left\{ t_n - \mathbf{m}_N^T \boldsymbol{\phi}(x_n) \right\}^2 \quad (N \gg M)$$

- These results are useful as they do not require computing the eigenspectrum of the Hessian.

# Another Example of Model Evidence

- We have seen in an earlier lecture that the conjugate prior for a Gaussian with unknown mean and unknown precision is a normal-gamma distribution.
- We can apply the same for the case of our likelihood

$$p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

for which the conjugate prior for  $\mathbf{w}$  and  $\beta$  is:

$$p(\mathbf{w}, \beta) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \mathcal{Gam}(\beta | a_0, b_0)$$

- It can be shown that the corresponding posterior for this takes the form:

$$p(\mathbf{w}, \beta | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \beta^{-1} \mathbf{S}_N) \mathcal{Gam}(\beta | a_N, b_N)$$

# Posterior Distribution

□ The posterior takes the form:

$$\begin{aligned} p(\mathbf{w}, \beta | \mathbf{t}) &\propto \beta^{N/2} \exp \left\{ -\frac{1}{2} \mathbf{w}^T \beta \Phi^T \Phi \mathbf{w} - \beta \mathbf{w}^T \Phi^T \mathbf{t} - \frac{1}{2} \beta \sum_{n=1}^N t_n^2 \right\} \\ &\quad \beta^{M/2} \exp \left\{ -\frac{1}{2} (\mathbf{w} - \mathbf{m}_0)^T \beta \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \right\} \beta^{a_0-1} e^{-b_0 \beta} \\ &\propto \beta^{N/2} e^{-\beta (\mathbf{w} - \mathbf{m}_N)^T \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N)} e^{\frac{1}{2} \mathbf{m}_N^T \beta \mathbf{S}_N^{-1} \mathbf{m}_N} e^{-\frac{1}{2} \beta \sum_{n=1}^N t_n^2} \beta^{M/2} e^{-\frac{1}{2} \mathbf{m}_0^T \beta \mathbf{S}_0^{-1} \mathbf{m}_0} \beta^{a_0-1} e^{-b_0 \beta} \\ &\propto \beta^{M/2} e^{-\beta (\mathbf{w} - \mathbf{m}_N)^T \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N)} \frac{1}{\Gamma(a_N)} b_N^{a_N} \beta^{a_N-1} e^{-b_N \beta} \\ &\propto \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \beta^{-1} \mathbf{S}_N) \mathcal{Gam}(\beta | a_N, b_N) \end{aligned}$$

where:

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N \left[ \mathbf{S}_0^{-1} \mathbf{m}_0 + \Phi^T \mathbf{t} \right], \quad \mathbf{S}_N^{-1} = \left[ \mathbf{S}_0^{-1} + \Phi^T \Phi \right] \\ a_N &= a_0 + \frac{N}{2}, \quad b_N = b_0 + \frac{1}{2} \left( \mathbf{m}_0^{-1} \mathbf{S}_0^{-1} \mathbf{m}_0 - \mathbf{m}_N^{-1} \mathbf{S}_N^{-1} \mathbf{m}_N + \sum_{n=1}^N t_n^2 \right) \end{aligned}$$



# Model Evidence

- The model evidence for our example is given as:

$$\begin{aligned} p(\mathbf{t}) &= \iint p(\mathbf{t} | \mathbf{w}, \beta) p(\mathbf{w} | \beta) d\mathbf{w} p(\beta) d\beta \\ &= \iint \left( \frac{\beta}{2\pi} \right)^{N/2} \exp \left\{ -\frac{\beta}{2} (\mathbf{t} - \Phi \mathbf{w})^T (\mathbf{t} - \Phi \mathbf{w}) \right\} \\ &\quad \left( \frac{\beta}{2\pi} \right)^{M/2} |\mathbf{S}_0|^{-1/2} \exp \left\{ -\frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \right\} d\mathbf{w} \\ &\quad \Gamma(a_0)^{-1} b_0^{a_0} \beta^{a_0-1} e^{-b_0\beta} d\beta = \\ &= \frac{b_0^{a_0}}{\left( (2\pi)^{M+N} |\mathbf{S}_0| \right)^{1/2}} \iint \exp \left\{ -\frac{\beta}{2} (\mathbf{t} - \Phi \mathbf{w})^T (\mathbf{t} - \Phi \mathbf{w}) \right\} \\ &\quad \exp \left\{ -\frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \right\} d\mathbf{w} \\ &\quad \Gamma(a_0)^{-1} \beta^{a_0-1} \beta^{N/2} \beta^{M/2} e^{-b_0\beta} d\beta \end{aligned}$$

# Model Evidence

- Using some of our [earlier results](#) in deriving the posterior:

$$p(\mathbf{t}) = \frac{b_0^{a_0}}{\left((2\pi)^{M+N} |\mathbf{S}_0|\right)^{1/2}} \iint \exp \left\{ -\frac{\beta}{2} (\mathbf{t} - \mathbf{m}_N)^T \mathbf{S}_N^{-1} (\mathbf{t} - \mathbf{m}_N) \right\} d\mathbf{w} \\ \exp \left\{ -\frac{\beta}{2} \left( \mathbf{t}^T \mathbf{t} + \mathbf{m}_0^T \mathbf{S}_0^{-1} \mathbf{m}_0 - \mathbf{m}_N^T \mathbf{S}_N^{-1} \mathbf{m}_N \right) \right\} \\ \Gamma(a_0)^{-1} \beta^{a_0-1} \beta^{M/2} \exp(-b_0 \beta) d\beta$$

- Performing the integration in  $\mathbf{w}$  and using the normalization factor for the Gamma distribution:

$$p(\mathbf{t}) = \frac{b_0^{a_0}}{\left((2\pi)^{M+N} |\mathbf{S}_0|\right)^{1/2}} (2\pi)^{M/2} |\mathbf{S}_N|^{1/2} \Gamma(a_0)^{-1} \underbrace{\int \beta^{a_0-1} \exp(-b_0 \beta) d\beta}_{\Gamma(a_0)/b_0^{a_0}} \\ = \frac{1}{(2\pi)^{N/2}} \frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \frac{b_0^{a_0}}{b_N^{a_0}} \frac{\Gamma(a_0)}{\Gamma(a_N)}$$



# Limitations of Fixed Basis Functions

- ❑ Up to now the basis functions  $f_j(\mathbf{x})$  are fixed before the training data set is observed.
- ❑ As a consequence, the number of basis functions grows exponentially with the dimensionality  $D$  of the input space.
- ❑ There are two properties of real data sets that we can exploit to alleviate the curse of dimensionality:
  - the data vectors  $\{\mathbf{x}_n\}$  typically lie close to a nonlinear manifold whose intrinsic dimensionality is smaller than that of the input space as a result of strong correlations between the input variables.
  - If we are using localized basis functions, we can arrange that they are scattered in input space only in regions containing data (radial basis functions, support vector and relevance vector machines).



# Adaptive Basis Functions

---

- ❑ Neural network models using adaptive basis functions having sigmoidal nonlinearities, can adapt the parameters so that the regions of input space over which the basis functions vary correspond to the data manifold.
- ❑ The target variables may have significant dependence on only a small number of possible directions within the data manifold.
- ❑ Neural networks can exploit this property by choosing the directions in input space to which the basis functions respond.



# Laplace Approximation

- As we have seen earlier, the Laplace approximation allows a Gaussian approximation of the parameter posterior about the maximum a posteriori (MAP) parameter estimate.
- Consider a data set  $\mathcal{D}$  and  $M$  models  $\mathcal{M}_i, i = 1, \dots, M$  with corresponding parameters  $\boldsymbol{\theta}_i, i = 1, \dots, M$ . We compare models using the posteriors:

$$p(\mathcal{M} | \mathcal{D}) \propto p(\mathcal{M})p(\mathcal{D} | \mathcal{M})$$

- For large sets of data  $\mathcal{D}$  (relative to the model parameters), the parameter posterior is approximately Gaussian around  $\boldsymbol{\theta}_m^{MAP}$  (can also use 2<sup>nd</sup> order Taylor expansion of the log-posterior):

$$p(\boldsymbol{\theta}_m | \mathcal{D}, M_m) \simeq (2\pi)^{-d/2} |\mathbf{A}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{\theta}_m - \boldsymbol{\theta}_m^{MAP})^T \mathbf{A}(\boldsymbol{\theta}_m - \boldsymbol{\theta}_m^{MAP})\right),$$
$$A_{ij} = -\frac{\partial^2 \log P(\boldsymbol{\theta}_m | \mathcal{D}, M_m)}{\partial \theta_{mi} \partial \theta_{mj}} \Big|_{\boldsymbol{\theta}_m^{MAP}}$$

# Laplace Approximation

- We can write the model evidence as

$$p(\mathcal{D} | \mathcal{M}_m) = \frac{p(\boldsymbol{\theta}_m, \mathcal{D} | \mathcal{M}_m)}{p(\boldsymbol{\theta}_m | \mathcal{D}, \mathcal{M}_m)}$$

- Using the Laplace approximation for the posterior of the parameters and evaluating the equation above at  $\boldsymbol{\theta}_m^{MAP}$ :

$$\begin{aligned} \log p(\mathcal{D} | M_m) &\simeq \log p(\boldsymbol{\theta}_m^{MAP}, \mathcal{D} | M_m) - \log p(\boldsymbol{\theta}_m^{MAP} | \mathcal{D}, M_m) \\ &\simeq \log p(\mathcal{D} | \boldsymbol{\theta}_m^{MAP}, M_m) + \log p(\boldsymbol{\theta}_m^{MAP} | M_m) + \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |A| \end{aligned}$$

- This Laplace approximation is used often for model comparison.
- Other approximations are also very useful:
  - Bayesian Information Criterion (BIC) (on the limit of  $N \rightarrow \infty$ )
  - MCMC (Sampling approach)
  - Variational Methods



# Bayesian Information Criterion

- We start with the Laplace approximation on the limit of large data sets  $N \rightarrow \infty$ ,

$$\log p(\mathcal{D} | M_m) \simeq \log p(\mathcal{D} | \boldsymbol{\theta}_m^{MAP}, M_m) + \log p(\boldsymbol{\theta}_m^{MAP} | M_m) + \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |A|$$

- As  $N$  grows,  $A$  grows as  $NA_0$  for some fixed matrix  $A_0$ , thus

$$\log |A| \rightarrow \log |NA_0| = \log(N^d |A_0|) = d \log N + \log(|A_0|) \xrightarrow{N \rightarrow \infty} d \log N$$

- Then the Laplace approximation is simplified as:

$$\log p(\mathcal{D} | M_m) \simeq \log p(\mathcal{D} | \boldsymbol{\theta}_m^{MAP}, M_m) - \frac{d}{2} \log N \quad (\text{limit } N \rightarrow \infty)$$

- Note interesting properties of (the easy to compute) BIC:

- No dependence on the prior
- One can use the MLE rather than the MAP estimate of  $\boldsymbol{\theta}_m$
- If not all parameters are well determined from the data,  
 $d$  = number of effective parameters.



# Another Example of Model Selection

- Let us consider a regression model with the following particulars ( $d = k + 1$  dimensional data):<sup>a</sup>

$$y \mid \mathbf{w}, \sigma^2, \Phi \sim \mathcal{N}_n(\Phi \mathbf{w}, \sigma^2 \mathbf{I}_n)$$

$$\mathbf{w} \mid \sigma^2, \Phi \sim \mathcal{N}_{k+1}(\mathbf{w}_0, \sigma^2 \mathbf{M}^{-1}), \mathbf{M} \text{ a } (k+1) \times (k+1) \text{ pos. def. symm. matrix}$$

$$\sigma^2 \mid \Phi \sim \text{InvGamma}(a, b), a, b > 0$$

$$\mathbf{M} = \mathbf{I}_{k+1} / c, c > 0 \text{ and } \mathbf{w}_0 = \mathbf{0}_{k+1}$$

- Our data are in a matrix form (dimension  $N \times (k + 1)$ ):

$$\Phi = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ 1 & x_{31} & x_{32} & \dots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

We only slightly change our notation here to conform with [the MatLab program](#) implementing this example (from Zoubin Ghahramani)





# Example: Likelihood Calculation

- We will derive the model evidence analytically. At first the likelihood can be written as:

$$\ell(\mathbf{w}, \sigma^2 \mid \mathbf{y}, \Phi) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \Phi\mathbf{w})^T (\mathbf{y} - \Phi\mathbf{w})\right)$$

- With simple algebra, we can rewrite the likelihood introducing the MLE estimate of the parameters as follows:

$$\begin{aligned} \ell(\mathbf{w}, \sigma^2 \mid \mathbf{y}, \Phi) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \Phi\mathbf{w}_{ML})^T (\mathbf{y} - \Phi\mathbf{w}_{ML}) - \frac{1}{2\sigma^2} (\mathbf{w} - \mathbf{w}_{ML})^T \Phi^T \Phi (\mathbf{w} - \mathbf{w}_{ML})\right) = \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} s^2 - \frac{1}{2\sigma^2} (\mathbf{w} - \mathbf{w}_{ML})^T \Phi^T \Phi (\mathbf{w} - \mathbf{w}_{ML})\right) \end{aligned}$$

where

$$\mathbf{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

$$s^2 \triangleq (\mathbf{y} - \Phi\mathbf{w}_{ML})^T (\mathbf{y} - \Phi\mathbf{w}_{ML})$$



# Computing Model Evidence

□ Our posterior is then of the following form:

$$p(\mathbf{w}, \sigma^2 / \mathbf{w}_{ML}, s^2, \Phi) = \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \sigma^{-k-n-2a-3} \exp\left(-\frac{1}{2\sigma^2} \left\{ s^2 + 2b + \mathbf{w}^T \left( \frac{1}{c} \mathbf{I} + \Phi^T \Phi \right) \mathbf{w} - 2\mathbf{w}^T \Phi^T \Phi \mathbf{w}_{ML} + \mathbf{w}_{ML}^T (\Phi^T \Phi) \mathbf{w}_{ML} \right\}\right)$$

□ The evidence is now computed as (use first the normalization of the *Inv Gamma* distribution):

$$\begin{aligned} p(y / \mathcal{M}) &= \iint \ell(\mathbf{w}, \sigma^2 | y, \Phi) p(\mathbf{w}, \sigma^2 / \mathbf{w}_{ML}, s^2, \Phi) d\mathbf{w} d\sigma^2 = \\ &= \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \iint \sigma^{-k-n-2a-3} \exp\left(-\frac{1}{2\sigma^2} \underbrace{\left\{ s^2 + 2b + \mathbf{w}^T \left( \frac{1}{c} \mathbf{I} + \Phi^T \Phi \right) \mathbf{w} - 2\mathbf{w}^T \Phi^T \Phi \mathbf{w}_{ML} + \mathbf{w}_{ML}^T (\Phi^T \Phi) \mathbf{w}_{ML} \right\}}_A\right) d\sigma^2 d\mathbf{w} = \\ &= \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \iint \sigma^{-k-n-2a-3} \exp\left(-\frac{1}{2\sigma^2} A\right) d\sigma^2 d\mathbf{w} = \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \iint (\sigma^2)^{(-k-n-2a-3)/2} \exp\left(-\frac{A/2}{\sigma^2}\right) d\sigma^2 d\mathbf{w} = \\ &= \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \iint (\sigma^2)^{\underbrace{-(k+n+2a+1)/2}_{\alpha}} \exp\left(-\underbrace{(A/2)}_{\beta} \frac{1}{\sigma^2}\right) d\sigma^2 d\mathbf{w} = \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \int \frac{\Gamma((k+n+2a+1)/2)}{(A/2)^{(k+n+2a+1)/2}} d\mathbf{w} = \\ &= \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \Gamma((k+n+2a+1)/2) 2^{(k+n+2a+1)/2} \int \left[ s^2 + 2b + \mathbf{w}^T \left( \frac{1}{c} \mathbf{I} + \Phi^T \Phi \right) \mathbf{w} - 2\mathbf{w}^T \Phi^T \Phi \mathbf{w}_{ML} + \mathbf{w}_{ML}^T (\Phi^T \Phi) \mathbf{w}_{ML} \right]^{-\frac{(k+n+2a+1)}{2}} d\mathbf{w} \end{aligned}$$

Useful formulas for the *Inv Gamma*:  $p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta}, \theta > 0$



# Computing Model Evidence

- The evidence can be further simplified (we will use now the normalization of the multivariate *Student-t* distribution)

$$\begin{aligned}
 p(\mathbf{y} / \mathcal{M}) &= \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \Gamma\left(\frac{d+n}{2} + a\right) 2^{\frac{d+n}{2}+a} \\
 &\int \left[ \underbrace{s^2 + 2b + \mathbf{w}_{ML}^T (\Phi^T \Phi) \mathbf{w}_{ML} - \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu}}_g + (\mathbf{w} - \boldsymbol{\mu})^T \left[ \underbrace{\left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right)^{-1}}_{\Sigma} \right] (\mathbf{w} - \boldsymbol{\mu}) \right]^{-\frac{(d+n)/2-a}{2}} d\mathbf{w} \\
 &\quad \mu = \left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right)^{-1} (\Phi^T \Phi) \mathbf{w}_{ML} \\
 &\quad \nu = n + 2a \\
 &\frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \Gamma\left(\frac{d+n}{2} + a\right) 2^{\frac{d+n}{2}+a} g^{-(d+n)/2-a} \int \left[ 1 + \frac{1}{g} \{(\mathbf{w} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{w} - \boldsymbol{\mu})\} \right]^{-\frac{(d+n+2a)/2}{2}} d\mathbf{w} = \\
 &\frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \Gamma\left(\frac{d+n}{2} + a\right) 2^{\frac{d+n}{2}+a} g^{-(d+n)/2-a} \int \left[ 1 + \frac{1}{\nu} \left\{ (\mathbf{w} - \boldsymbol{\mu})^T \left( \frac{g}{n+2a} \Sigma \right)^{-1} (\mathbf{w} - \boldsymbol{\mu}) \right\} \right]^{-\frac{(d+\nu)/2}{2}} d\mathbf{w}
 \end{aligned}$$

Useful formulas for the *Student-t*:

$$\begin{aligned}
 p(\boldsymbol{\theta}) &= \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2) \nu^{d/2} \pi^{d/2}} |\Sigma|^{-1/2} \\
 &\times \left( 1 + \frac{1}{\nu} (\boldsymbol{\theta} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) \right)^{-(\nu+d)/2}
 \end{aligned}$$



# Computing Model Evidence

□ Performing the integration in  $\beta$  in the last slide:

$$\begin{aligned}
 p(\mathbf{y} / \mathcal{M}) &= \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} 2^{\frac{d+n}{2}+a} g^{-(d+n)/2-a} \Gamma\left(\frac{n}{2} + a\right) (n+2a)^{d/2} \pi^{d/2} \left(\frac{g}{n+2a}\right)^{d/2} \left|\left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right)\right|^{-1/2} = \\
 &\frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} c^{d/2} \Gamma\left(\frac{n}{2} + a\right) \left(\frac{1}{2} s^2 + b + \frac{1}{2} \mathbf{w}_{ML}^T (\Phi^T \Phi) \mathbf{w}_{ML} - \frac{1}{2} \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu}\right)^{-n/2-a} \left|\left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right)\right|^{-1/2} = \\
 &\frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} c^{d/2} \Gamma\left(\frac{n}{2} + a\right) \underbrace{\left(\frac{1}{2} (\mathbf{y} - \Phi \mathbf{w}_{ML})^T (\mathbf{y} - \Phi \mathbf{w}_{ML}) + b + \frac{1}{2} \mathbf{w}_{ML}^T (\Phi^T \Phi) \mathbf{w}_{ML} - \frac{1}{2} \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu}\right)}_{\mathcal{E}}^{-n/2-a} \left|\left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right)\right|^{-1/2}
 \end{aligned}$$

□ Using some of the earlier definitions,

$$\mathbf{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} \quad s^2 \triangleq (\mathbf{y} - \Phi \mathbf{w}_{ML})^T (\mathbf{y} - \Phi \mathbf{w}_{ML}) \quad \boldsymbol{\mu} = \left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right)^{-1} (\Phi^T \Phi) \mathbf{w}_{ML}$$

we can simplify  $\mathcal{E}$  as:

$$\begin{aligned}
 \mathcal{E} &= \frac{1}{2} (\mathbf{y} - \Phi \mathbf{w}_{ML})^T (\mathbf{y} - \Phi \mathbf{w}_{ML}) + b + \frac{1}{2} \mathbf{w}_{ML}^T (\Phi^T \Phi) \mathbf{w}_{ML} - \frac{1}{2} \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} = b + \frac{1}{2} \mathbf{y}^T \mathbf{y} + \frac{1}{2} \mathbf{w}_{ML}^T (\Phi^T \Phi) \mathbf{w}_{ML} - \mathbf{y}^T \Phi \mathbf{w}_{ML} + \frac{1}{2} \mathbf{w}_{ML}^T (\Phi^T \Phi) \mathbf{w}_{ML} \\
 &- \frac{1}{2} \mathbf{w}_{ML}^T (\Phi^T \Phi) \left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right)^{-1} \left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right) \left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right)^{-1} (\Phi^T \Phi) \mathbf{w}_{ML} = b + \frac{1}{2} \mathbf{y}^T \mathbf{y} + \mathbf{y}^T \Phi (\Phi^T \Phi)^{-1} (\Phi^T \Phi) (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} \\
 &- \mathbf{y}^T \Phi (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} - \frac{1}{2} \mathbf{w}_{ML}^T (\Phi^T \Phi) \left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right)^{-1} (\Phi^T \Phi) (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} = b + \frac{1}{2} \mathbf{y}^T \mathbf{y} - \frac{1}{2} \mathbf{y}^T \Phi \left(\frac{1}{c} \mathbf{I} + \Phi^T \Phi\right)^{-1} \Phi^T \mathbf{y}
 \end{aligned}$$



# Model Evidence

- The final evidence in analytical form is given as:

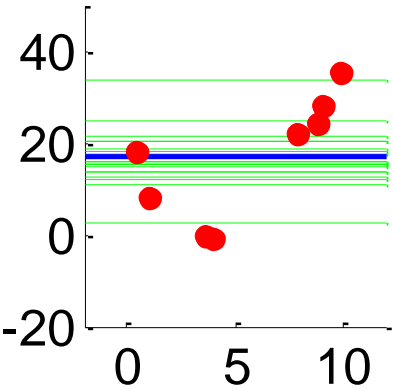
$$p(\mathbf{y} / \mathcal{M}) = \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} c^{d/2} \Gamma\left(\frac{n}{2} + a\right) \left( b + \frac{1}{2} \mathbf{y}^T \mathbf{y} - \frac{1}{2} \mathbf{y}^T \Phi \left( \frac{1}{c} \mathbf{I} + \Phi^T \Phi \right)^{-1} \Phi^T \mathbf{y} \right)^{-n/2-a} \left| \left( \frac{1}{c} \mathbf{I} + \Phi^T \Phi \right) \right|^{-1/2}$$

- Compare this with what is given in this [MatLab Implementation](#).
- The model evidence and samples of different order ( $M$ ) regression models are given below. The specific data of the problem can be found in the MatLab file.
- We are looking for the order of the polynomial that maximizes the evidence. Note that the MatLab implementation utilized random input/output data.

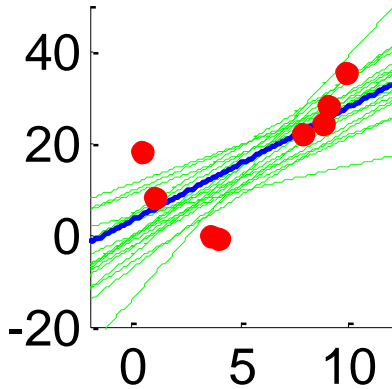


# Bayesian Model Comparison

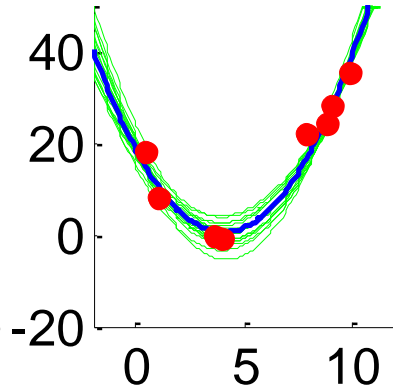
$M = 0$



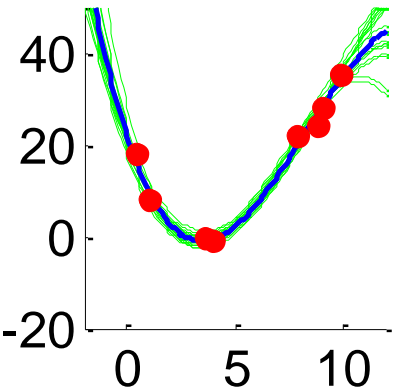
$M = 1$



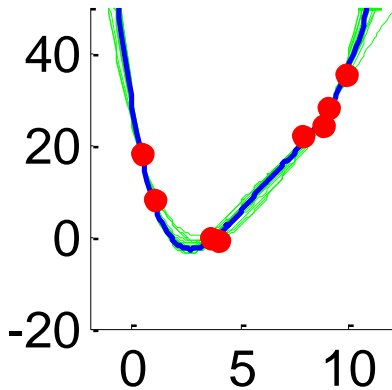
$M = 2$



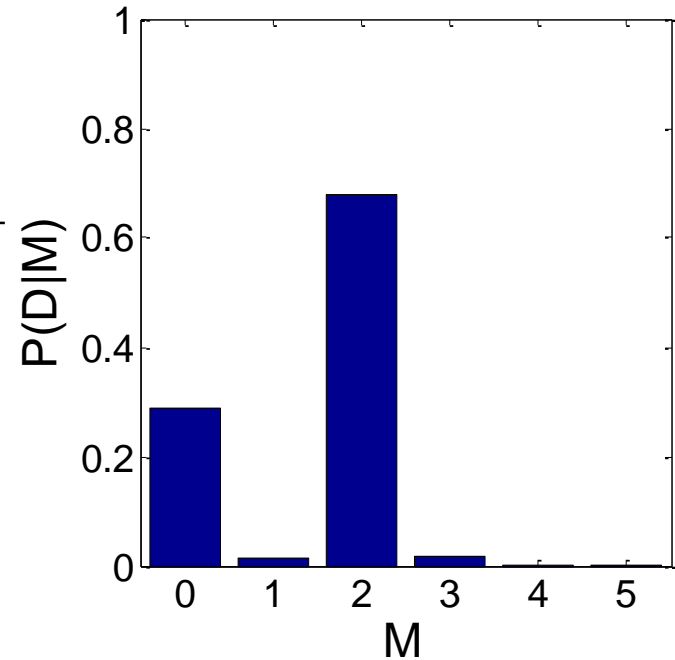
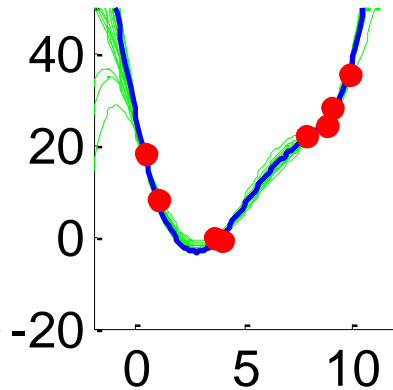
$M = 3$



$M = 4$



$M = 5$



[MatLab implementation](#) of Bayesian Model Selection (from [Zoubin Ghahramani](#))

